Center for Statistik
Handelshøjskolen i København

MPAS
Tue Tjur
November 2007

$\mathbf{R}^2$–størrelser for logistiske regressionsmodeller.

Indledning.

I en lineær normalfordelingsmodel, f.eks. en mulitipel regressionsmodel af formen

$$y_i \sim N\left(\alpha + \beta x_i + \ldots, \sigma^2\right)$$

defineres som bekendt de *fittede værdier* $\hat{\mu}_i$ som de estimerede middelværdier af observationerne

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i + \ldots$$

og *residualerne* er størrelserne $y_i - \hat{\mu}_i$. *Residualkvadratsummen* er kvadratsummen af residualerne,

$$\mathrm{SSD}_{\mathrm{res}} = \sum (y_i - \hat{\mu}_i)^2$$

og ved *modelkvadratsummen* vil vi i dette notat forstå størrelsen

$$\mathrm{SSD}_{\mathrm{mod}} = \sum (\hat{\mu}_i - \bar{y})^2 \,.$$

Under antagelse af, at modellen indeholder et konstantled — hvilket helt generelt er en forudsætning for at tale om R^2–størrelser — gælder det, at gennemsnittet $\bar{y}$ af observationerne er det samme som gennemsnittet af de fittede værdier, så $\mathrm{SSD}_{\mathrm{mod}}$ er simpelthen kvadratsummen af de fittede værdiers afvigelser fra deres gennemsnit. Desuden gælder det, at vektoren af fittede værdier er ortogonal på vektoren af residualer, hvoraf følger (af "Pytagoras") at

$$\mathrm{SSD}_{\mathrm{y}} = \sum (y_i - \bar{y})^2 = \mathrm{SSD}_{\mathrm{mod}} + \mathrm{SSD}_{\mathrm{res}} \,.$$

Ved modellens *forklaringsgrad*, eller *determinationskoefficienten*, forstås størrelsen

$$R^2 = \frac{\mathrm{SSD}_{\mathrm{mod}}}{\mathrm{SSD}_{\mathrm{y}}} = 1 - \frac{\mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_{\mathrm{y}}} \,.$$

Denne størrelse fortolkes som den andel af den samlede kvadratafvigelsessum $\mathrm{SSD}_{\mathrm{y}}$, der er opfanget af modellen, idet resten ($\mathrm{SSD}_{\mathrm{res}}/\mathrm{SSD}_{\mathrm{y}}$) er den andel som skyldes tilfældig variation.

Forklaringsgraden er ikke en teststørrelse, og det er heller ikke en størrelse som siger noget om hvorvidt modellen holder eller ej. R^2 er et rent

deskriptivt mål for hvor god en modellen er til at forudsige observationerne, sammenlignet med den simple model der antager at observationerne er identisk fordelte. For store datasæt vil man ofte være interesseret i at simplificere modellen ved at fjerne f.eks. vekselvirkningsled af høj orden, også selv om de er signifikante. Hvis dette ikke ændrer R^2 ret meget kan man sige, at dette er acceptabelt ud fra en væsentlighedsvurdering. Ligeledes, hvis man sammenligner modeller som ikke umiddelbart kan sammenlignes ved testning, for eksempel varianter af "samme" model med baggrundsvariable der er transformerede på forskellige måder, kan det være god mening i at vælge den model, der har den største forklaringsgrad (eller den mindste residualkvadratsum, det kommer ud på et i dette tilfælde).

Forklaringsgraden for en logistisk regressionsmodel.

Betragt en logistisk regressionsmodel

$$p_i = P(y_i = 1) = \frac{\exp(\alpha + \beta x_i + \dots)}{1 + \exp(\alpha + \beta x_i + \dots)}$$

hvor $y_1, \dots, y_n$ betegner de uafhængige, binære responser. Med $\hat{p}_i$ betegner vi de fittede værdier, altså de estimerede sandsynligheder for "succes" $y_i = 1$. I det følgende antages at modellen har et konstantled (som α ovenfor), i modsat fald giver det følgende ikke mening.

Der findes i litteraturen temmelig mange definitioner af, hvad man skal forstå ved forklaringsgraden for en logistisk regressionsmodel. Vi skal ikke omtale dem allesammen, men til gengæld tilføje en hjemmelavet variant, samt gøre rede for at nogle af disse definitioner fører til approksimativt samme værdi af R^2.

Når man ser på definitionen fra de lineære modeller er der to forslag som ligger lige for:

$$R_1^2 = \frac{\sum(\hat{p}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

og

$$R_2^2 = 1 - \frac{\sum(y_i - \hat{p}_i)^2}{\sum(y_i - \bar{y})^2}$$

På grund af antagelsen om, at modellen har et konstantled, gælder det at summen af de fittede værdier er lig med summen af selve observationerne (denne identitet er nemlig en af likelihoodligningerne, den man får ved differentiation efter α). Heraf følger at gennemsnittet $\bar{y}$ af observationerne er det samme som gennemsnittet $\bar{\hat{p}}$ af de fittede værdier. Så kvadratsummen $\sum(\hat{p}_i - \bar{y})^2$ kan også fortolkes som de fittede værdiers kvadratiske variation omkring deres eget gennemsnit. Derimod gælder det *ikke*, at vektoren $(\hat{p}_i - \bar{y})$ er vinkelret på residualvektoren $(y_i - \hat{p}_i)$, så til forskel fra det lineære tilfælde vil R_1^2 og R_2^2 i almindelighed være

forskellige. Dog er de, som vi nu skal se, approksimativt ens. "Approksimativt" skal her forstås sådan, at vi vil antage at antallet af observationer er så stort, at de estimerede sandsynligheder $\hat{p}_i$ uden nævneværdig fejl kan erstattes med de tilsvarende sande successandsynligheder p_i, og de summer som indgår i udtrykkene i det følgende uden nævneværdig (relativ) fejl kan erstattes med deres middelværdier med henvisning til de store tals lov.

I det følgende lader vi q_i betegne de komplementære sandsynligheder $1-p_i$. De tre kvadratsummer, som indgår i tæller og nævner i udtrykkene for R_1^2 og R_2^2 ovenfor, kan nu approksimeres med funktioner af de ukendte parametre på følgende måde.

Tælleren i udtrykket for R_1^2:

$$\sum(\hat{p}_i - \bar{y})^2 \approx \sum(p_i - \bar{p})^2 \,.$$

Tælleren i udtrykket for R_2^2:

$$\sum(y_i - \hat{p}_i)^2 \approx \sum(y_i - p_i)^2 = \sum y_i^2 + \sum p_i^2 - 2\sum y_i p_i$$

$$= \sum y_i + \sum p_i^2 - 2\sum y_i p_i \approx \sum p_i + \sum p_i^2 - 2\sum p_i^2$$

$$= \sum p_i - \sum p_i^2 = \sum p_i q_i \,.$$

Den fælles nævner:

$$\sum(y_i - \bar{y})^2 \approx \sum(y_i - \bar{p})^2 = \sum y_i^2 + n\bar{p}^2 - 2\bar{p}\sum y_i$$

$$\approx \sum y_i + n\bar{p}^2 - 2\bar{p}(n\bar{p}) \approx n\bar{p} - n\bar{p}^2 = n\bar{p}(1-\bar{p}) = n\bar{p}\,\bar{q} \,.$$

Følgende udregning viser, at disse approksimationer til de tre kvadratsummer opfylder den relation, som svarer til relationen $\mathrm{SSD_{mod}} + \mathrm{SSD_{res}} = \mathrm{SSD_y}$ i det lineære tilfælde:

$$\sum(p_i - \bar{p})^2 + \sum p_i q_i = \sum\left(p_i^2 + \bar{p}^2 - 2p_i\bar{p} + p_i(1-p_i)\right)$$

$$= \sum\left(p_i^2 + \bar{p}^2 - 2p_i\bar{p} + p_i - p_i^2\right) = \sum p_i^2 + n\bar{p}^2 - 2n\bar{p}^2 + n\bar{p} - \sum p_i^2$$

$$= n\bar{p} - n\bar{p}^2 = n\bar{p}(1-\bar{p}) = n\bar{p}\,\bar{q} \,.$$

Heraf følger, at

$$R_1^2 \approx \frac{\sum(p_i - \bar{p})^2}{n\bar{p}\,\bar{q}} = 1 - \frac{\sum p_i q_i}{n\bar{p}\,\bar{q}} \approx R_2^2 \,.$$

I praksis er det således formodentlig ligegyldigt, om vi vælger R_1^2 eller R_2^2 som udtryk for forklaringsgraden.

Man kunne her få den tanke, at det i en eller anden forstand er "bedre" at bruge den fælles approksimation af R_1^2 og R_2^2 med $\hat{p}_i$'erne indsat i stedet for p_i' erne. Men det fører ikke til noget nyt, idet der simpelthen gælder $\sum(y_i - \bar{y})^2 = n\bar{y}(1-\bar{y}) = n\bar{\hat{p}}\,\bar{\hat{q}}$, og dermed

$$R_1^2 = \frac{\sum(\hat{p}_i - \bar{\hat{p}})^2}{n\bar{\hat{p}}\,\bar{\hat{q}}}\,.$$

Et nyt forslag.

I det følgende vil vi referere til en simpel kontroltegning, som man i almindelighed bør lave i forbindelse med logistisk regression, nemlig de to "parallelle" histogrammer som viser fordelingen af de fittede værdier $\hat{p}_i$ for henholdsvis "successer" ($y_i = 1$) og "fiaskoer" ($y_i = 0$). Grupperingen kan passende vælges så enhedsintervallet opdeles i 10 lige store intervaller, for store datasæt eventuelt 20. På denne tegning kan man blandt andet se, om modelspecifikationen holder (en slags grafisk variant af Hosmer–Lemeshows test). Søjlehøjderne i de to histogrammer skal jo, på nær tilfældig støj, have et bestemt forhold. Ser vi for eksempel på de observationer i der har $\hat{p}_i \approx 0.25$ — altså dem der ligger mellem 0.2 og 0.3 — skulle der gerne være ca. 3 (=0.75/0.25) gange så mange fiaskoer som successer, dvs. søjlen i "fiasko–histogrammet" skal være ca. tre gange så høj som den tilsvarende søjle i "succes–histogrammet".

Tegningen siger desuden noget om forklaringsgrad. Hvis modellen er god til at forudsige, vil successernes $\hat{p}_i$'er klumpe sig sammen tæt på 1 i histogrammets højre side, medens fiaskoernes klumper sig sammen omkring 0 i venstre side. Omvendt er en model som er dårlig til at forudsige de binære udfald karakteriseret ved, at $\hat{p}_i$'erne klumper sig sammen et sted inde i midten, svarende til at alle successandsynlighederne er næsten ens.

De to mål for forklaringsgrad, som vi har set på, kan aflæses af denne tegning (underforstået at vi kan gøre inddelingen så fin som vi har lyst til). R_2^2 måler jo direkte, om successernes $\hat{p}_i$'er ligger tæt på 1 og fiaskoernes tæt på 0, og nævneren normerer netop på en sådan måde at vi får forklaringsgrad 1 hvis $\hat{p}_i = y_i$ for alle i. R_1^2 måler simpelthen $\hat{p}_i$'ernes varians i forhold til den maksimalt mulige varians, som er selve observationernes varians (må man gå ud fra, det er ikke umiddelbart indlysende hvordan man skal bevise det). At variansen i tælleren beregnes for *alle* observationer, uden hensyn til hvilke der er successer og hvilke der er fiaskoer, og alligevel fører til approksimativt samme resultat som R_2^2, kan måske virke lidt underligt. Men forklaringen er naturligvis, at det gennemgående er de store $\hat{p}_i$'er der hører til successerne og de små som hører til fiaskoerne. Hvis bare variansen er tæt på observationernes varians *og* søjlerne i histogrammerne ellers har de forhold, som de skal have

ifølge modellen, så får vi netop en stærkt venstreforskudt fordeling for fiaskoerne og en stærkt højreforskudt fordeling for successerne.

Med udgangspunkt i den fortolkning af forklaringsgrad, som knytter sig til kontroltegningen, kan man hævde at et endnu mere nærliggende forslag til et mål for forklaringsgraden er differensen mellem middelværdierne i de to fordelinger. Den varierer jo netop mellem 0 og 1, med 1 svarende til en rent deterministisk model ($\hat{p}_i = y_i$) og 0 svarende til at alle $\hat{p}_i$'erne er ens. Vi foreslår altså

$$R_3^2 = \bar{\hat{p}_1} - \bar{\hat{p}_0}$$

hvor $\bar{\hat{p}_1}$ betegner gennemsnittet af fittede værdier for successerne, og $\bar{\hat{p}_0}$ tilsvarende gennemsnittet af de fittede værdier for fiaskoerne. Denne størrelse har umiddelbart ikke megen lighed hverken med R_1^2 eller R_2^2. Men det viser sig at den er approksimativt lig med de to andre, idet den simpelthen er **eksakt** lig med deres gennemsnit. Dette ses af følgende udregninger. Med betegnelserne

$n_1 = \sum y_i = n\bar{y}$ = antal successer, og

$n_0 = \sum(1 - y_i) = n(1 - \bar{y})$ = antal fiaskoer

kan R_3^2 omskrives således,

$$R_3^2 = \frac{\sum y_i\hat{p}_i}{n_1} - \frac{\sum(1-y_i)\hat{p}_i}{n_0} = \frac{n_0\sum y_i\hat{p}_i - n_1\sum(1-y_i)\hat{p}_i}{n_1 n_0}$$

$$= \frac{n\sum y_i\hat{p}_i - n_1\sum\hat{p}_i}{n^2\bar{y}(1-\bar{y})} = \frac{\sum y_i\hat{p}_i - \bar{y}\sum\hat{p}_i}{n\bar{y}(1-\bar{y})} = \frac{\sum\hat{p}_i(y_i-\bar{y})}{n\bar{y}(1-\bar{y})},$$

medens gennemsnittet af de to tidligere forslag er

$$\frac{1}{2}\left(R_1^2 + R_2^2\right) = \frac{1}{2}\left(\frac{\sum(\hat{p}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2} + 1 - \frac{\sum(y_i-\hat{p}_i)^2}{\sum(y_i-\bar{y})^2}\right)$$

$$= \frac{1}{2}\left(1 + \frac{\sum(\hat{p}_i-\bar{y})^2 - \sum(y_i-\hat{p}_i)^2}{\sum(y_i-\bar{y})^2}\right)$$

$$= \frac{1}{2}\left(1 + \frac{\sum\hat{p}_i^2 + \sum\bar{y}^2 - 2\sum\hat{p}_i\bar{y} - \sum y_i^2 - \sum\hat{p}_i^2 + 2\sum y_i\hat{p}_i}{\sum y_i^2 - n\bar{y}^2}\right)$$

$$= \frac{1}{2}\left(1 + \frac{\sum\bar{y}^2 - 2\sum\hat{p}_i\bar{y} - \sum y_i + 2\sum y_i\hat{p}_i}{\sum y_i - n\bar{y}^2}\right)$$

$$= \frac{1}{2}\left(1 + \frac{n(\bar{y}^2-\bar{y}) + 2\sum(y_i-\bar{y})\hat{p}_i}{\sum y_i - n\bar{y}^2}\right)$$

$$= \frac{1}{2}\left(1 + \frac{-n\bar{y}(1-\bar{y}) + 2\sum(y_i - \bar{y})\hat{p}_i}{n\bar{y}(1-\bar{y})}\right)$$

$$= \frac{1}{2}\left(1 - 1 + \frac{2\sum(y_i - \bar{y})\hat{p}_i}{n\bar{y}(1-\bar{y})}\right) = \frac{\sum(y_i - \bar{y})\hat{p}_i}{n\bar{y}(1-\bar{y})} = R_3^2\,.$$

Det står os således frit for at vælge R_3^2 som det relevante mål for forklaringsgrad, og på grund af denne størrelses simple intuitive fortolkning forekommer dette at være et rimeligt valg.

ROC kurven og arealet under den.

ROC står for **R**eciever **O**perating **C**haracteristic og har i virkeligheden noget med kalibrering af en radar at gøre. Så betegnelsen er ikke særligt velvalgt, men sådan hedder det altså, og det dækker over følgende.

For et givet $p_0 \in]0,1[$ kan man konstruere en slags "predikteret respons" som

$$y_i^* = \begin{cases} 1 \text{ hvis } \hat{p}_i > p_0 \\ 0 \text{ hvis } \hat{p}_i \leq p_0 \end{cases}$$

Betragt den 2×2–tabel, der klassificerer alle observationerne efter værdien af den faktiske respons y_i (=0 eller 1, tabellens rækker) og den predikterede respons y_i^* (=0 eller 1, tabellens søjler). Denne tabel siger noget om hvor god modellen er til at prediktere. Hvis der er store tal i diagonalen og små tal i de to andre celler, er modellen åbenbart god til at skelne successer fra fiaskoer på basis af de fittede værdier. Disse tabeller kan aflæses af den tidligere omtalte kontroltegning (del histogrammerne i to ved hjælp af den lodrette linie $p = p_0$, antallene i tabellen kan så aflæse som antal til venstre/højre for linien i de to histogrammer.

Lad $n_1(p_0)$ betegne antallet af observationer med $y_i = 1$ og $\hat{p}_i > p_0$ og lad $n_0(p_0)$ tilsvarende betegne antallet af observationer med $y_i = 0$ og $\hat{p}_i > p_0$. ROC kurven er den kurve der udgøres af (eller fås ved at forbinde) de n punkter

$$(n_0(p_0), n_1(p_0))\ ,\ \ p_0 \in]0,1[$$

i et koordinatsystem, hvor den vandrette akse løber fra 0 til n_0 og den lodrette fra 0 til n_1. Idet man dog normalt bagefter omskalerer akserne så de løber fra 0 til 1 (eller 0% til 100%).

En perfekt model, der forudsiger 100% korrekt for en eller anden værdi af p_0 er karakteriseret ved at kurven først løber fra (0,0) lodret op til (0,1), og derefter vandret hen til (1,1). En dårlig model vil, må man gå ud fra, sortere observationernes fittede værdier helt uafhængigt af de faktiske responser, hvilket (p.g.a. de store tals lov) svarer til at kurven følger diagonalen.

På baggrund af denne fortolkning har man som en alternativ definition af forklaringsgraden foreslået

$$R_4^2 = \text{ arealet under ROC kurven.}$$

Værdien $R_4^2 = 1$ svarer således til en perfekt model, medens modeller med $R_4^2 \leq 0.5$ må anses for at være helt ubrugelige.

Det følger af ovenstående forklaring, at denne størrelse kan fortolkes på følgende måde: Lad $\hat{p}_S$ være en fittet værdi, trukket tilfældigt blandt de n_1 fittede værdier for successerne, og lad tilsvarende $\hat{p}_F$ være trukket tilfældigt blandt de n_0 fittede værdier for fiaskoerne. Så er

$$R_4^2 = P\left(\hat{p}_S \geq \hat{p}_F\right)$$

(eller $P\left(\hat{p}_S > \hat{p}_F\right)$, afhængigt af hvad vi forstår ved arealet under ROC kurven).